RAND

Teaching Practices and Student Achievement: Evaluating Classroom-Based Education Reforms

Laura S. Hamilton, Daniel F. McCaffrey, Stephen P. Klein, Brian M. Stecher, Abby Robyn, Delia Bugliari

DRU-2603-EDU

July 2001

RAND Education

The RAND unrestricted draft series is intended to transmit preliminary results of RAND research. Unrestricted drafts have not been formally reviewed or edited. The views and conclusions expressed are tentative. A draft should not be cited or quoted without permission of the author, unless the preface grants such permission.

RAND is a nonprofit institution that helps improve policy and decisionmaking through research and analysis. RAND's publications and drafts do not necessarily reflect the opinions or policies of its research sponsors.

Running Head: TEACHING PRACTICES AND STUDENT ACHIEVEMENT

Teaching Practices and Student Achievement:

Evaluating Classroom-Based Education Reforms

Laura S. Hamilton

Daniel F. McCaffrey

Stephen P. Klein

Brian M. Stecher

Abby Robyn

Delia Bugliari

RAND

This research was supported by the National Science Foundation, Division of Elementary, Secondary, and Informal Education, under grant ESI-96-15809.

Abstract

A number of recent efforts to improve mathematics and science instruction have focused on professional development activities designed to promote instruction that is consistent with professional standards such as those published by the National Council of Teachers of Mathematics. The National Science Foundation's Systemic Initiatives (SI) program is an example. We gathered data from 11 SI programs and investigated relationships between student achievement and the teachers' use of standards-based instruction. We used multiple measures of student achievement, and pooled results across the 11 sites using a planned meta-analytic approach. We observed small but consistent positive relationships between teachers' use of standards-based instruction and student achievement, but we were unable to detect any differences across types of achievement measures. Our data and results suggest a number of limitations of methods that are currently used for implementing and evaluating reforms, and provide some lessons for future evaluations of reforms that promote instructional change.

Teaching Practices and Student Achievement: Evaluating Classroom-Based Education Reforms

Many education reform efforts seek to improve student learning through changes in classroom practices. These reforms include structural changes, such as reduction of class size, as well as specific changes in curriculum or teaching practices. The federal government and other agencies have invested significant resources in evaluations of these programs. The evaluations are often intended to inform future decisions about funding as well as to justify resources spent on particular programs. Central to most of these evaluations is the use of student achievement test scores as a measure of outcomes to determine whether the reforms are producing the intended effects on students. Evaluators face a number of challenges, including teachers who feel overburdened and who believe their students spend too much time taking tests. In this paper we describe our approach to studying one set of reforms and discuss some of the challenges inherent in these kinds of evaluations.

The National Science Foundation's Systemic Initiatives (SI) program provides an example of a reform that seeks to influence instructional practices. States and districts receive SI funds to implement math and science reforms that include efforts to change the way these subjects are taught. Although the SI programs are intended to address the entire system surrounding math and science education, their presumed effects on student achievement depend in large part on effective implementation in the classroom. Substantial resources have been invested in the program, but the link between classroom practices and student achievement has not been established. NSF therefore funded a study of this relationship in a number of sites where the reforms were being implemented.

Our approach had three major features that distinguish it from most educational program evaluations: (1) We designed a planned meta-analysis, with 11 separate sites and similar data from each; (2) we used existing student achievement data but supplemented it with additional measures wherever possible in an effort to have multiple measures of student achievement; and (3) we examined within-school variation in implementation rather than simply comparing participating and non-participating schools. The evaluation revealed small relationships between student achievement and teachers' use of classroom practices promoted by the SI's in most sites. More important, it suggests a number of limitations of methods that are currently used for implementing and evaluating reforms, and provides some lessons for future evaluations of reforms that promote instructional change.

The paper begins with background information on the SI programs. We summarize existing evidence on the effectiveness of these efforts and the difficulties researchers face in measuring relevant student outcomes. We then describe our approach to studying the problem, including our samples, measures, and methods of analysis, followed by the results. The conclusion of this paper provides a summary of our major findings, a discussion of the strengths and limitations of our approach, and directions for future research and evaluation.

Background on Systemic Reform

A cornerstone of the systemic reform initiatives is the alignment among all parts of the educational system, including curriculum, instruction, assessment, teacher preparation, and state and local policies such as graduation requirements. Such alignment is perceived as necessary for promoting change in the classroom and, ultimately, improving student performance (Smith & O'Day, 1991). Systemic reform efforts have resulted in part from observations that addressing one component of the educational system tended to be ineffective due to constraints imposed by other parts of the system (Hill, 1995; Knapp, 1997).

NSF's programs included the Statewide Systemic Initiatives (SSI), Urban Systemic Initiatives (USI), Rural Systemic Initiatives (RSI), and Local Systemic Change (LSC) programs. They were intended to provide resources to promote system-wide change, and many were funded at a level of several million dollars over multiple years. The USI program, for example, funded 20 large urban districts with awards of up to $15 million for a five-year period, focusing on cities where high proportions of children live in poverty. The program is described as a "comprehensive and systemic effort to stimulate fundamental, sweeping, and sustained improvement in the quality and level of K-12 science, mathematics, and technology (SMT) education" (Williams, 1998, p. 7). Taken together, these Systemic Initiative (SI) programs received approximately $100 million per year in NSF funding during the 1990's. In addition, most sites supplemented their NSF grants with additional local contributions. For example, sites have used Title I funds, corporate donations, and grants from private foundations to support and expand their systemic initiatives (Williams, 1998).

Although these programs vary in scope and emphasis, all are relatively long-term (five years, with a small number of SSIs being extended for an additional five years), and all attempt coordinated reform, aligning various parts of the educational system with one another. These initiatives, in theory at least, generally involve the development of ambitious curriculum and performance standards and the mobilization of all components of the system to support and enable all student to reach those standards (Consortium for Policy Research in Education, 1995).

To be effective, these reforms must ultimately be adopted by teachers and take hold in the classroom (Tyack & Cuban, 1995). Thus, a primary emphasis of the systemic reform initiatives involves promotion of teaching practices that are assumed to facilitate student learning. Most initiatives offer professional development to teachers, and this component constitutes a fairly large proportion of the budget. For example, the SSI sites spent nearly one third of their first-year budgets on in-service training for teachers, more than on any other category of spending (Shields, Corcoran, & Zucker, 1994). The goal of most of this training is to increase teachers' use of classroom practices that are believed to improve achievement.

The kinds of practices being promoted by NSF, and by numerous other agencies and reformers, are consistent with curriculum standards and guidelines that have been published by the National Research Council (1996), the American Association for the Advancement of Science (1993), and the National Council of Teachers of Mathematics (1989). Common to all of these documents is an emphasis on instruction that engages students as active participants in their own learning and that enhances the development of complex cognitive skills and processes. Specific practices that are endorsed include cooperative learning groups, inquiry-based activities, use of materials and manipulatives, and open-ended assessment techniques. All of these practices are intended to support active rather than passive learning, to promote the application of critical thinking skills, and to provide opportunities to apply math and science learning to real-world contexts.

Earlier Evaluations of SI Programs

Most SI sites worked with outside organizations to evaluate their efforts, and additional external evaluations have been commissioned by NSF. Although these evaluations typically focused more on implementation (e.g., type and frequency of professional development offered to teachers, level of participation among teachers) than on outcomes, many of the SI sites have reported improvement in student test scores (Williams, 1998). However, most of these reports offer little if any evidence to tie this improvement directly to SI participation.

A large-scale study of SSI programs conducted by SRI International revealed small but statistically significant differences in test scores that favored participating over non-participating schools in four of seven sites (Laguarda, 1998). Although these results are encouraging, factors other than SI participation may be driving the differences. First, the analyses did not control for pre-existing differences in the teachers and students found in SSI and non-SSI schools. Sites often implement large-scale reforms in phases, and those schools that participate in the earlier phases may differ in important ways from those that participate in later phases. Second, variations in implementation within schools were not considered. The fact that a school is considered part of the reform effort does not guarantee that all the teachers in the school are responding in the intended manner. Other researchers have found that teachers' use of reform practices is influenced by many factors, including the nature and frequency of professional development participation (Cohen & Hill, 1998; Weiss, Montgomery, Ridgway, & Bond, 1998) and the degree to which they understand the subject matter (Cohen & Ball, 1990). Finally, the data were collected and analyzed by site personnel rather than by the external evaluators, and there may be variations in quality of that data and analytic approaches used across sites.

The absence of comprehensive evaluation data on SI programs has led some policymakers to express skepticism about the value of these programs (Fox, 1998). Others have called for more rigorous evaluations that focus on student achievement and relate it to the degree of implementation of the reform. There is some evidence of a positive relationship between the practices promoted by the systemic initiatives and student achievement in mathematics and science, and we review this evidence in the next section.

Evidence of Relationships Between Teaching Practices and Student Achievement

If the systemic initiatives do result in improved student achievement, it is undoubtedly due in large part to what occurs in the classroom. For this reason, professional development and the promotion of good instructional practices are critical to the success of the initiatives. Research provides some evidence of the effectiveness of some of the individual practices endorsed by the reforms. An experiment conducted by Ginsburg-Block and Fantuzzo (1998), for example, showed that low-achieving elementary students who were assigned to problem solving or peer collaboration conditions obtained higher math scores and reported higher levels of motivation than did students who received neither of these interventions. Several other studies have demonstrated the value of peer tutoring and collaboration (e.g., Fantuzzo, King, & Heller, 1992; Greenwood, Carta, & Hall, 1988; Webb & Palincsar, 1996), as well as the benefits of contextualizing instruction in real-world problems (Verschaffel & De Corte, 1997).

A few studies have focused on relationships between student achievement and teachers' use of combinations of these practices. Cohen and Hill (1998) studied teacher-reported use of several practices consistent with the 1992 California Mathematics Framework and found that frequency of use was positively related to scores on the California Learning Assessment System (CLAS) mathematics test at the school level, after controlling for demographic characteristics. The set of teaching practices examined in this study was similar to those being advocated and supported by the systemic initiatives. Mayer (1998) found small positive or null relationships between a similar set of practices and student scores on a standardized multiple-choice test. Thus there is some evidence that, in certain contexts at least, use of reform practices is related to higher student achievement.

Measuring Student Achievement

One difficulty in conducting evaluations of ongoing programs in general, and of the systemic initiatives in particular, is a lack of appropriate measures of student achievement. Most state testing programs, for example, rely heavily or exclusively on multiple-choice items (Education Week, 2000), a format that does not always lend itself to measuring many of the scientific inquiry and mathematical problem solving skills encouraged by the systemic initiatives. Science reforms are particularly difficult to evaluate because not all states administer science assessments, and those that do often limit them to a few grade levels (Goertz & Duffy, 2001).

State testing programs also typically fail to provide data that can be used to track the progress of individual students over time. Many states, for example, test students only in a few grades (e.g., 4th, 8th, and 10th; see Goertz & Duffy, 2001). This forces evaluators to focus on changes in scores of successive cohorts of students, which confound the effects of reforms with differences among the groups of students. In addition, improvements in scores over time, which are often cited as evidence of beneficial effects of reforms on student learning, may in many cases reflect inappropriate narrowing of the curriculum or teaching to the test (Koretz & Barron, 1998; Linn, 2000). This problem is especially likely when the tests are part of a high-stakes accountability system or when the same form of a test is administered multiple times. For all of these reasons, it is desirable to supplement existing state tests with additional measures whenever possible.

Goals of the Mosaic Project

The study described in this report was designed to examine the relationship between teaching practices and student achievement in mathematics and science, a relationship which is at the heart of the systemic initiatives. We focused on a measures of the extent to which teachers engaged in activities consistent with the systemic reforms (using a scale we call "reform practices", described below). We gathered data from a variety of SI sites and used multiple measures of achievement, producing a "mosaic" of evidence about this relationship—hence the name of the study, the "Mosaic" project. We examined this relationship at the level of the classroom, which allowed us to address variations in degree of implementation both within and across schools. In addition, we measured student achievement using both multiple-choice and open-response tests, including some hands-on science tasks that we developed and administered ourselves. Finally, we used information on student demographics and prior achievement to control for pre-existing differences among students. The Mosaic study was conducted in eleven sites, which are described below. An earlier report (citation deleted for anonymity) presents results for the first six sites, and provides additional details on methodology.

It is important to recognize that this project was not a comprehensive evaluation of the systemic reform initiatives. These initiatives are multi-faceted, multi-year efforts to bring about changes in classroom practice and in other aspects of the educational system. The reform sites have adopted a wide range of strategies to recruit and train teachers in new methods, to implement new curricula, to provide appropriate materials, to encourage and sustain change at the school level, and to instill greater interest in mathematics and science. Their success at these tasks is the subject of a comprehensive evaluation undertaken by SRI International (e.g., Corcoran, Shields, and Zucker, 1998; Shields, Marsh, & Adelman, 1998). Our study, by contrast, focused on one key aspect of the SI reforms, the relationship between instructional practices and student outcomes.

Methods

We collected data from eleven sites—six during the spring of 1997 and six during the spring of 1998 (one site participated both years). Our specific procedures for site selection, subject and grade level selection, and data collection are described in the following sections.

Site, School, Subject, and Grade Level Selection

Because we knew that it would be difficult to study the relationship between reform-based instructional practices and achievement in the absence of a reasonable degree of reform, we selected sites in a way that maximized the probability of encountering substantial numbers of teachers using reform practices. NSF staff proposed sites in which there were indications that large numbers of teachers had adopted the reform practices in their classroom. NSF drew its recommendations from site visits and from progress reports submitted by the grantees.

School district and program personnel at each site specified the grade level(s) and subject(s) in which they believed reform practices were most pervasive, and then nominated schools to participate in the study. The same basic research design was used at each site. We asked local staff to select approximately ten schools in which there was good reason to believe mathematics and/or science reforms had been implemented, and ten demographically similar schools in which the reforms had yet to be implemented. All of the sites had been involved in the reform for more than one year, but had yet to implement the reforms in all schools in the site. We used the nominations only to ensure variation in teaching practices; we did not compare the high- and low-implementing schools with one another directly. Table 1 lists the grade(s) and subject(s) of data collection and the numbers of teachers and students participating at each site. Between 85 and 100 percent of selected schools participated in the study, and within sites, teacher survey completion rates ranged from 71 to 100 percent, with most sites achieving close to 100 percent participation.

Table 1

Sites, Grades, Subjects, Numbers of Participants, and Assessments

Site	Grade	Subject(s)	Number of Teachers	Number of Students	Tests	Added for Mosaic Study?	Prior Year Test Scores Available?
1	3[a]	Math	46	804	State multiple-choice math	No	No
					State open-response math	No	
2	5	Math	100	1651-1686	State multiple-choice math	No	Yes
					Commercial open-ended math	Yes	
		Science	99	1639-1662	Commercial multiple-choice science	Yes	
					Hands-on science[b]	Yes	
3	5	Math	73	1366-1451	Commercial multiple-choice math	No	Yes
					Commercial open-ended math	Yes	
		Science	74	1367-1438	Standards-based multiple-choice science[c]	No	
					Standards-based open-ended science[c]	No	
4	5	Science	45	909-932	Standards-based multiple-choice science[c]	No	Yes
					Standards-based open-ended science[c]	No	
5	7	Math	48	2937-3018	State multiple-choice math	No	No
					Commercial open-ended math	Yes	
		Science	33	2047-2079	Commercial multiple-choice science	Yes	
					Hands-on science[b]	Yes	
6	7	Math	57	3237	Commercial multiple-choice math[d]	No	Yes
		Science	52	3279	Commercial multiple-choice science[d]	No	
7	7	Math	57	3127-3145	Commercial multiple-choice math	No	Yes
					Commercial open-ended math	Yes	
		Science	52	2870	Commercial multiple-choice science	No	
8	5	Science	37	1637-1641	Commercial multiple-choice science	No	Yes
					State open-ended science	No	
9	4	Science	116	1783-1786	Commercial multiple-choice science	Yes	No
					Commercial open-ended science	Yes	
10	4	Math	76	1244-1248	State multiple-choice math	No	No
					State open-ended math	No	
		Science	76	1265-1270	State multiple-choice science	No	
					State open-ended science	No	
11	8	Math	28	1163	State multiple-choice math	No	No
					State open-ended math	No	
		Science	18	1033	State multiple-choice science	No	
					State open-ended science	No	
12	5	Math	67	1507-1592	Commercial multiple-choice math	No	Yes
					State multiple-choice math	No	
					State open-response math	No	

[a] At this site, we studied teaching practices for third-grade teachers and measured the relationships with student test scores that were gathered during the following fall when students had advanced to the fourth grade.
[b] See (*reference deleted for anonymity*) for a description of tasks and scoring guides.
[c] This test was developed by a consortium of educators and researchers, and was designed to be aligned with NSF-supported reform efforts.
[d] In this site, we were unable to schedule any open-ended testing.

Student Data

We obtained student scores on the mathematics and science assessments regularly administered at each site, and supplemented these with additional assessments, where feasible, to provide both multiple-choice and open-response scores. Supplementary tests were chosen in consultation with local staff, who were encouraged to select measures that they believe were reasonably well aligned with their reform efforts. Hands-on science tasks developed by (*institution deleted for anonymity*) were made available, and some sites opted to use them. Mosaic project staff trained exercise administrators to administer some of the supplementary measures, including the hands-on tasks. All other tests were administered by the classroom teachers or by test administrators who worked at the local sites. Table 1 indicates the types of tests administered in each site.

In all but one site, students completed a standardized multiple-choice assessment in mathematics and/or science depending on the site designation, and an open-response test that required students to produce, rather than select, their responses. One site administered only multiple-choice tests, and we were unable to schedule additional testing due to time constraints. We used existing tests wherever possible, including state-developed tests and commercially available standardized tests. The column “Added for Mosaic Study?” in Table 1 indicates whether we supplemented the district or state's testing program with additional measures or relied only on those measures already used by the sites. Later we discuss some of the advantages and limitations of using state and district test scores for the purposes of evaluation.

To control for pre-existing differences in student achievement, we obtained district or state test scores in the relevant subject from the spring prior to our data collection. In most sites, prior year test scores were missing for 5-10 percent of the student sample. We used hierarchical Bayesian models (Schafer, 1997) to impute multiple values for each missing value. The imputation models included all variables used in our regression models as well as contemporaneous reading and math scores. The models also accounted for the hierarchical structure of the data with students nested within classrooms.

In five sites we were unable to obtain prior year test scores because the state or district did not administer tests in the relevant grade or did not maintain individual student records. In these cases we used contemporaneous reading and language scores (i.e., scores on a reading test that was administered at approximately the same time as the tests we used as outcome measures) as covariates. Both prior year and contemporaneous scores serve as measures of student achievement. Unlike prior year scores, however, contemporaneous scores are not necessarily independent of the instructional practices measures by our surveys: If instruction in math or science involves activities that promote the use of verbal skills, for example, this instruction could improve reading or language scores. Including contemporaneous scores as covariates could absorb some of the effects of instruction and result in an incorrectly estimated relationship between practices and achievement in math or science. The alternative approach, which would involve excluding an achievement covariate altogether, is also problematic. We conducted sensitivity analyses and determined that including contemporaneous scores was the most appropriate approach, resulting in a very slight attenuation of the relationship between reform practices and achievement (for details on these analyses see *(reference deleted for anonymity)*). The last column in Table 1 indicates for which sites we had prior year test scores.

Finally, we obtained demographic data, which in most sites included race/ethnicity, gender, participation in free or reduced-price lunch programs, language background, and participation in special education or gifted programs. These data were used to verify that comparison schools were similar to implementing schools on student demographics, enrollment, and grade span, and were included as covariates in the analysis of relationships between teaching practices and student achievement. These data also enabled us to study whether these relationships varied as a function of student characteristics.

The inability to link data from students and teachers is a factor that hinders many evaluations of instructional programs. Few district or state data systems maintain these links in a readily usable form, especially at the secondary grades when students often have a different teacher for each subject. Most of our sites lacked these links, so we collected class rosters from teachers and used these to make the links.

Teacher Data

Our primary measure of teaching practices in each site was a questionnaire developed by Horizon Research, Inc. (HRI). This instrument is a modified version of a questionnaire that has been validated and used extensively by HRI to evaluate the implementation of the Local Systemic Change (LSC) initiatives. Questionnaires were administered to all teachers in a school teaching the targeted subject and grade level.

Typically, the site coordinator or assistant distributed the questionnaires either individually or at after-school meetings and then collected completed questionnaires in individual, sealed envelopes for return to us.

We created separate questionnaires for mathematics and science teachers, but many of the items were identical across subjects. Questions asked teachers to report the frequency of various instructional practices ranging from traditional (e.g., "have students watch me [teacher] do a science demonstration") to reform ("[students] conduct investigations where they develop their own procedures for addressing a question or problem"). General topics included: amount of time spent on science/mathematics; approach to introducing a new topic; typical teacher instructional practices; typical student activities; types of written assignments; and methods of assessing student learning.

Although NSF did not mandate a particular curriculum or a specific set of teaching strategies for the Systemic Initiatives, there was an emerging consensus among math and science educators about what should be taught and how it should be presented. (National Council of Teachers of Mathematics, 1989; National Research Council, 1996). In light of this consensus, it is not surprising that the systemic reform programs adopted very similar content and instructional goals. An independent evaluation of the SSI program reported that "across the states there was remarkable similarity in the perceived shortcomings of current practices and the set of desirable reforms in curriculum content and instructional strategies." (Shields, Marsh, & Adelman, 1998; page 2). The shared content goals included greater emphasis on conceptual understanding of math and science concepts, the application of this knowledge to everyday situations, and the integration of concepts across subjects. The instructional emphasis was equally distinct. Rather than viewing students as passive learners who absorb unrelated facts and procedures, the reforms sought to engage students actively in their own learning, to be sensitive to each student's learning style, to increase the use to technology, and to utilize new forms of assessment for instructional planning. In mathematics this meant more "data gathering and analysis, statistics, geometry and visualization, discovery learning and 'constructivist' approaches;" in science more "scientific processes, such as observation, comparison, experimentation, hypothesis generation, hypothesis-testing, and theory building" (New Jersey SSI Proposal, p. 7; quoted in Shields, March & Adelman, 1998, p.3). Our measures of instructional strategies were designed to be consistent with this espoused commonality of purpose.

In addition, teachers completed a brief demographic section, providing information about their college degree, teaching certification, coursework in mathematics and/or science, gender, ethnicity, and years of teaching experience. In sites where science or mathematics specialists delivered instruction instead of the regular classroom teacher, we administered surveys to the specialists and also asked the respondent to clarify the teaching situation.

Analysis

The primary purpose of this study was to investigate the degree to which student achievement was associated with teachers' use of reform-based instructional practices. There are numerous approaches to modeling data with a possible intra-class correlation that can result from sampling multiple students within the same classroom. In this paper we use ordinary least squares to estimate the regression coefficients and use a nonparametric estimator of the standard error of these estimated coefficients. This estimator adjusts the standard errors to account for possible correlation among responses from students with the same teacher (McCaffrey & Bell, 1997). Unlike standard hierarchical linear models, it is robust to assumptions about the correlation among scores for students from the same classroom. Because the focus of our study is on the relationship between teaching practices and student outcomes, and not the correlation among students within a classroom, we used the nonparametric standard error estimates.

We found that teacher background variables did not provide any additional explanatory power and therefore we do not include them in the results reported here. At each site, we conducted separate analyses for mathematics and science and for open-response and multiple-choice tests. We fit these models using individual student data, with all students from the same classroom receiving the same values for the teacher-reported instructional practices measure.

The use of data from multiple sites provides an opportunity to conduct a planned meta-analysis. We therefore also conducted pooled analyses which combined data from all six sites to produce a single estimate of the coefficient relating teaching practices to student achievement. This approach provides results that are similar to what would be obtained by pooling individual scores and fitting a random coefficients model with interactions between sites and the covariates, but it permits the pooling of data across sites without requiring identical models in every site (Goldstein, 1995). We conducted separate analyses by subject (math or science) and test format (multiple-choice or open-response).

Results

We first present summaries of teachers' reported use of instructional practices. We then present our findings with regard to the relationships between use of these practices and student achievement in each site. Finally, we describe results of an analysis of differences between open-response and multiple-choice achievement measures.

Distributions of Teaching Practices

Based on exploratory factor analyses of the questionnaire items, we identified two clusters of items and created scales from these by simply summing the scores on each item. The first scale measured the teacher's use of "reform practices." Each of the 22 items in this scale asked teachers to report the frequency of use of a particular reform practice (such as cooperative groups, portfolios, hands-on activities, and extended investigations). We also created a 5-item "traditional practices" scale based on items that measured the amount of time teachers spent in traditional teaching practices (such as textbook work, lectures, and short-answer tests). The score for each teacher was simply the average item response across items. All items used a 5-point Likert scale, so teachers' scores could range from 1 (rarely or never using any of the practices) to 5 (engaging in all practices daily or almost daily). Across sites and subjects, the average alpha coefficient was 0.92 for reform practices and 0.70 for traditional practices[1]. This distinction between reform-related practices and more traditional practices is consistent with the kinds of definitions used in other research on math and science reform (e.g., Cohen & Hill, 1998; Smerdon, Burkam, & Lee, 1999).

It is important to note that the two scales are not opposites of one another. A principal components analysis of the questionnaire data resulted in the identification of these two scales in each site, suggesting that teachers may use both reform and traditional practices to different degrees. Correlations between the two scales ranged across sites from moderately negative to moderately positive, with many close to zero. It is possible for teachers to be high on both scales because the scale scores do not indicate the total amount of time spent on these practices, but rather the frequency with which they are used. Thus a teacher who intersperses lecture-style teaching with opportunities for student discussion in every lesson might score high on both scales. In addition, there are other activities that are not addressed by either scale, so it is possible for teachers to receive low scores on both.

In each site we found a wide range of practices on both the reform and the traditional scales. Table 2 provides descriptive information for each combination of site and subject (math or science). There are some differences across sites in the score ranges and variability. Although these differences could influence the likelihood of detecting relationships with achievement, the results we discuss in later sections show no clear patterns with respect to these differences. Inspection of the distributions of scores suggests that range restriction is not a problem in any of our sites.

Table 2

Descriptive Statistics on Teaching Practices Scales, by Site

Site	Subject	Scale	Mean	St. Dev.	Minimum	Maximum
1	Math	Reform	3.20	.56	1.82	4.50
2	Math	Reform	3.61	.58	2.05	4.64
3	Math	Reform	3.38	.56	1.68	4.68
5	Math	Reform	3.01	.59	2.00	4.64
6	Math	Reform	3.34	.59	1.50	4.73
7	Math	Reform	3.32	.54	2.00	4.50
10	Math	Reform	3.79	.83	1.60	4.90
11	Math	Reform	3.25	.97	1.60	4.60
12	Math	Reform	3.63	.45	2.76	4.88
1	Math	Traditional	3.73	.66	2.00	5.00
2	Math	Traditional	3.63	.66	2.00	4.80
3	Math	Traditional	3.33	.57	2.20	5.00
5	Math	Traditional	3.40	.65	1.80	4.80
6	Math	Traditional	3.73	.63	2.20	5.00
7	Math	Traditional	3.80	.51	2.60	5.00
10	Math	Traditional	3.25	1.08	1.00	5.00
11	Math	Traditional	3.07	1.06	1.40	4.60
12	Math	Traditional	4.08	.56	2.40	5.00
2	Science	Reform	3.27	.68	1.00	4.55
3	Science	Reform	3.33	.64	1.64	4.41
4	Science	Reform	3.57	.53	1.95	4.36
5	Science	Reform	3.22	.69	1.64	4.53
6	Science	Reform	3.28	.69	1.45	5.00
7	Science	Reform	3.31	.58	1.55	4.27
8	Science	Reform	3.44	.48	2.56	4.81
9	Science	Reform	3.00	.58	1.00	4.19
10	Science	Reform	3.66	.85	1.80	4.90
11	Science	Reform	3.33	.77	1.80	4.60
2	Science	Traditional	3.34	.75	1.00	5.00
3	Science	Traditional	2.86	.60	2.20	4.40
4	Science	Traditional	2.65	.70	1.20	4.00
5	Science	Traditional	3.78	.67	2.60	4.90
6	Science	Traditional	3.62	.59	2.00	5.00
7	Science	Traditional	3.66	.58	2.20	4.80
8	Science	Traditional	3.31	.63	1.60	4.40
9	Science	Traditional	2.43	.54	1.00	4.00
10	Science	Traditional	2.72	1.02	1.00	5.00
11	Science	Traditional	3.26	.79	2.20	4.60

Note: All scores are averages across items on the 5-point scale described in the text of the article.

Although it is not shown in this table, we observed large variability within schools, regardless of whether they were originally classified by site staff as high- or low-implementing. High-implementing schools were likely to include at least one teacher who reported infrequent use of reform practices, and low-implementing schools often had teachers who reported using reform practices liberally. This underscores the importance of linking student outcomes directly to his or her teacher rather than to a school-wide average.

Relationships Between Teaching Practices and Student Achievement

As indicated earlier, we examined relationships between instructional practices and student achievement using regression models that controlled for prior achievement and student background characteristics. We estimated separate models for the reform and traditional scales for each of the four subject-by-test-format combinations (math multiple-choice, math open-response, science multiple-choice, and science open-response). Below we provide detailed results only for the reform practices scale, partly because it is more directly relevant to NSF's approach, but also because the five-item traditional scale did not always function well in our analyses: There were significant nonlinearities for several sites, making it difficult to pool results across sites, and the standard errors for the traditional practices coefficients tended to be large, due in part to the relatively low reliability of the 5-item scale[2]. When we pooled results across models for which a linear term was appropriate, in none of the four pooled analyses was the coefficient for traditional practices significantly different from zero.

Figures 1 through 4 (see pages 40-43) provide an overview of our reform practices analyses in each site, as well as the pooled results across sites. The relationships depicted in the figures are the estimated coefficients from our regression models for the reform practices scale. We report standardized coefficients, which represent the expected difference in test score standard deviation units for a one standard deviation unit increase in scores on the reform scale. The dark dot represents the point estimate for the coefficient and the gray bar represents 95 percent confidence interval for that point estimate. The bottom bar in each figure shows the estimated coefficient from the pooled analysis, described later.

Figure 1, which shows relationships between use of reform practices and achievement on open-response math tests, indicates that in seven of the eight sites where we had open-response mathematics tests, higher test scores were associated with greater use of reform practices (i.e., the estimated coefficient was positive). However, the coefficients were statistically significantly greater than zero in only four of these sites. Similarly, Figure 2 shows that for almost all of the participating sites, higher multiple-choice test scores in mathematics were associated with greater use of reform practices, although only one of the estimates was statistically significantly different from zero.

Figures 3 and 4 show that greater use of reform practices in science was associated with higher test scores on both open-response and multiple-choice measures in science. Again most of the estimated coefficients were extremely small and were not statistically significantly different from zero, even though an inspection of coefficients across sites show a consistent pattern of a weak positive relationship between the reform practice scale and test scores.

As shown in Figures 1 through 4, the relationship between reform practices and test scores is at most small in almost all our models. For example, one of the larger coefficients was 0.09, an estimate of the relationship between reform practices and open-response science tests in Site 2 (see Figure 3). In this site, our model suggests that for a teacher using all of the reform practices monthly, the average student was predicted to score at about the 48th percentile in the site on the test, while for a teacher using all of the reform practices weekly we would predict that a similar student would score at about the 54th percentile[3]. Smaller changes in percentiles would be expected in most of the other sites. Compared with the coefficients for most of the student background characteristics (e.g., an average coefficient of 0.54 across sites for participation in free- or reduced-price lunch programs), all of the relationships we observed may be considered small.

Open-response measures are often perceived as more appropriate indicators of student achievement in the context of standards-based teaching than are multiple-choice measures, in part because they appear to tap skills that are similar to those that are emphasized in the classroom (e.g., open-ended problem solving). Inspection of the regression coefficients suggests that there may be a difference in the strength of the relationships between instructional practices and the two types of test, but it is small. Later we discuss a test of the statistical significance of this difference.

The bottom bars in Figures 1 through 4 show the pooled estimates of the standardized regression coefficients for each of the four analyses. The coefficients and confidence interval bounds are also presented in Table 3. The confidence intervals exclude zero in all four cases, though the lower bounds are very close to zero. Nevertheless, taking all of our data into account, we find small, positive relationships between reform-based instruction and student achievement in math and science, measured by both multiple-choice and open-response tests.

Table 3

Standardized Regression Coefficients for Reform Practices, Pooled Across Sites

Subject	Test Format	Weighted Average Coefficient	Lower Bound of 95% Confidence Interval	Upper Bound of 95% Confidence Interval
Math	OR	0.051	0.017	0.085
Math	MC	0.028	0.003	0.053
Science	OR	0.054	0.025	0.083
Science	MC	0.037	0.017	0.057

Differences Between Test Formats

Consistent with the individual site results, inspection of the coefficients from the pooled analyses suggested slightly larger relationships between open-response scores and reform teaching practices than between multiple-choice scores and reform teaching practices, especially in math. This finding is consistent with the hypothesis that the former type of test is more closely aligned with the reforms and therefore better able to detect effects. To test the statistical significance of this difference, we calculated the difference in standard deviation units between each student's score on the open-response test and his or her score on the multiple-choice test in the same subject. We then modeled these differences as a function of teaching practices and student background covariates. The analysis was repeated for both subjects and for all sites.

In very few individual sites was the difference between formats statistically significant, nor were the differences significant when we pooled results across sites. Table 4 presents the coefficients from the pooled analysis, along with 95% confidence intervals. The coefficient for math is 0.031. This implies that across sites, the expected increase in student math scores for a unit increase in a teacher's score on the reform scale was 0.032 standard deviation units higher for open-response tests than for multiple-choice tests. However our estimate was not statistically significantly different from zero. The estimate for science was even smaller. In addition, we found a relatively large between-site variance in these estimated differences, even after controlling for sampling error within site. In other words, we found that the difference in the sensitivity of open-response and multiple-choice tests varied from site to site. This variation is to be expected, given the large variations in test type within a format (e.g., open-ended science tests included both hands-on and paper-and-pencil, short-answer measures).

Table 4

Pooled Estimates of Differences between Formats

Subject	Weighted Average Coefficient	Lower Bound of 95% Confidence Interval	Upper Bound of 95% Confidence Interval
Math	0.031	-0.001	0.064
Science	0.010	-0.023	0.042

Thus, although inspection of regression coefficients suggests that open-response tests functioned differently from multiple-choice tests, our data do not provide sufficient evidence to support the claim that the formats differ in their sensitivity to the effects of the reform. Even so, the consistency in the patterns we observed, and that fact that educators involved in these reforms often assert that open-response tests are generally more closely aligned with their efforts, suggest that further investigation of format differences is appropriate and warranted. As states continue to develop standards-based assessments, and as results from these assessments are increasingly used in evaluations of educational programs as well as for high-stakes accountability purposes, it is critical that the instructional sensitivity of different test formats be examined.

Limitations of our Approach

There are several caveats that need to be considered when interpreting the results of this study. As with most educational research, our inability to investigate effects using an experimental design limits the inferences that can be made from results. Perhaps the primary problem is that without random assignment of students and teachers to treatments, we cannot be certain that the relationships we observed can be attributed solely to classroom practices. There may be other differences in student characteristics across classrooms that contribute to differences in performance and that influence what teachers do. For example, teachers may tend to engage in more reform-based practices with higher achieving students, or may simply be more highly skilled teachers than those who do not engage in these practices. Controlling for prior achievement and examining relationships with teacher background, as we have done here, is helpful but does not eliminate the problem completely.

A second limitation is the lack of information on what led teachers to utilize particular practices. Some may have adopted certain strategies as a result of participation in the professional development activities that are provided by the SI funds, but there are many other potential sources which would be difficult to capture with a paper-and-pencil questionnaire. The large variability in teaching practices within schools, which was observed for SI as well as non-SI schools, suggests that factors other than SI participation are influencing teachers' decisions about how to teach. Our initial intent was not to determine the reasons for teachers' use of practices, but information on this would be helpful to those who are designing and implementing professional development programs.

A third weakness of our approach stems from the use of questionnaires to measure instructional practices. Like any such measure, our items are subject to inaccurate responses, particularly those that reflect social desirability. Perhaps more importantly, our questions addressed only the frequency with which teachers used particular practices and did not address the ways in which they were used or the overall quality of the instruction. Clearly, some approaches to using cooperative groups are more effective and more consistent with the intent of the reform than others, but we cannot detect these differences using our questionnaires. Multiple classroom observations, interviews, and inspection of classroom materials would undoubtedly provide a better measure of instructional practice. This type of data, however, is considerably more expensive to collect, and is usually only done on a small scale. Our questionnaire items are similar to those that have been used in numerous evaluations of this type, and to those that have been administered as part of some national longitudinal surveys. There is clearly a need for cost-effective, valid measures of instructional practices. We are currently involved in a project that is intended to examine paper-and-pencil alternatives to traditional methods of measuring classroom practices, and we hope that the products of this project will be useful in future evaluations.

Discussion

As illustrated by Figures 1-4, the relationships between student achievement and teachers' use of instructional practices supported by the SI reforms tend to be positive but small, particularly in comparison to relationships between achievement and student background characteristics such as socioeconomic status and ethnicity. If, in fact, the observed relationships represent the effects of teaching practices on student achievement, their small magnitude may not be surprising given the brief period of time (less than one academic year) which was captured by teachers' questionnaire responses. Use of particular instructional strategies in a single course during a single school year would not be expected to lead to effects as large as those associated with student background characteristics. Several years of exposure may be needed to achieve a reasonably large effect. This suggests the need for longitudinal investigations, discussed below.

The direction of relationships was fairly consistent across sites, but their magnitudes displayed some variation. There are several potential sources of this variation. First, our models differed slightly across sites because we relied on locally available data to construct covariates. Second, various aspects of SI program implementation, such as the amount and quality of professional development activities, undoubtedly affected the kinds of teaching practices that were used. Even if two teachers report using reform practices with similar frequency, their approaches to those practices may differ substantially and may reflect specific features of the local reform program. Third, the achievement measures used in each site varied on a number of dimensions, including psychometric quality (e.g., reliability), content, and degree of alignment with the local curriculum.

This last source of differences has implications for future evaluations of SI's and other reforms. Most evaluations rely on locally available student achievement data, in large part because it is expensive and often not feasible to administer additional measures

Many principals and teachers believe that their students spend far too much time taking the tests that are required by the district and/or state, and are therefore reluctant to volunteer for supplementary testing. Locally developed and administered tests may also be preferred because they are presumed to be more closely aligned with local curriculum standards than would a measure chosen and administered by outside evaluators. In many of our sites, however, test development lagged far behind the reform implementation, leaving local personnel to rely on tests that they did not necessarily believe were ideal measures of student learning. It is likely that most large-scale evaluations will have to continue to use tests that are not fully adequate for their purposes, and to find ways to combine information from a diverse set of outcome measures.

Although the overall differences we observed between multiple-choice and open-response tests were not significant, the general pattern suggests that format effects should be investigated further. In particular, it raises questions concerning whether the two types of tests measure different constructs. Most advocates of systemic reform believe that traditional, multiple-choice tests do not adequately reflect the range of competencies that the reforms are expected to develop, and that tests requiring students to construct their answers and to engage in complex problem-solving are more appropriate. Our results do not indicate that this is necessarily the case, but the question deserves further investigation, particularly given the resources that many states and districts are devoting to open-ended testing.

Challenges for Program Evaluation

Countless education reform evaluations are currently underway, and many of these focus on programs that try to change what happens in the classroom. The federal government recently funded a number of evaluations of comprehensive school reform models, for example, in an effort to determine which models show the most promise for improving student achievement. Most of these models include guidelines for curriculum and instruction, and many espouse instructional approaches similar to those we have studied in the context of NSF's Systemic Initiatives programs.

To obtain a clear picture of how these reforms have been implemented in the classroom, and to understand the process by which they influence student learning, it is important to examine variations in instructional practice both within and among schools that vary in degree of reform adoption, and to link the degree of implementation in the classroom to student outcomes. We tried to do that here by using the three approaches that we identified at the beginning of this paper: (1) a cross-site analysis that allowed us to replicate the study in a variety of contexts; (2) multiple outcome measures with controls for prior achievement; and (3) measurement of implementation at the classroom rather than school level, and direct links between teachers and students. Although our approach generated some informative results, the discussion of limitations above provides some suggestions for improving the way this type of evaluation is carried out.

First, although the planned meta-analysis approach is promising, it suffers from some of the drawbacks of traditional meta-analyses—most significantly, cross-study variations in how constructs are measured and how models are specified. Although we were able to control these variations in ways that are not possible with traditional meta-analyses, limitations in data systems (e.g., missing covariates) and lack of a common outcome measure prevented us from specifying identical models across sites. Many states are currently engaged in efforts to build better data systems, and these will certainly enhance future evaluation efforts. The resources needed to create a data system are substantial, but the investment will undoubtedly result in more useful information that can guide school reform efforts.

A related issue is the need to link student outcomes to the specific classroom environment to which the student was exposed. We were able to do this by collecting rosters, but this was an enormously time-consuming activity, and would not be feasible in a very large evaluation such as those that include representative samples of schools from multiple states. Building teacher-student links into large-scale data systems would dramatically expand the kinds of questions evaluators could ask, and would result in more refined information on what approaches are effective. Although all sorts of practical and political issues make this an extremely difficult task, it has been done in some states (e.g., Tennessee) and is worth exploring further.

Effective evaluation of educational programs also requires an appropriate outcome measure. Almost all evaluations rely on state and district test scores as indicators of student achievement and learning. This is the least costly and burdensome approach, but it may not provide the most accurate information on program effects. Our data suggest that estimated relationships between instruction and achievement may differ when achievement is measured in a closed-ended versus an open-ended way. Currently the majority of state tests rely heavily or exclusively on multiple-choice items, and the use of the more expensive open-response format is likely to decline further as states increase the numbers of grades and subjects tested (and therefore the cost of testing). In addition, in states with high-stakes testing programs, gains on state tests are often not replicated on other tests (e.g., Koretz & Barron, 1998). In a few of our sites we observed unusual relationships among state test scores and scores on the tests we administered, suggesting that at minimum the two sets of tests are capturing different aspects of student achievement. In short, the method used for measuring achievement matters, but most evaluations do not have sufficient funding to permit the use of multiple outcome measures. Evaluators need to look carefully at the measures that they are using, and explore alternative methods for refining the information (e.g., the use of subsets of items that may be especially closely aligned with a particular curriculum reform).

Finally, assessment and evaluation should be built into reform programs from the outset. In particular, we reiterate others' calls for an increased emphasis on randomized experiments in education. Despite advances in statistical modeling, it is not possible to account fully for student and teacher characteristics that may be confounded with a program's effects. Better measures of student and teacher background are helpful, but cannot capture all differences. As new reforms are implemented in small numbers of schools, wherever feasible the schools should be chosen in a way that minimizes pre-existing differences. In addition, ongoing student assessment that is aligned with the program should be an integral part of the program package, so that the data necessary to evaluate the program's effects are collected from the very beginning. This would enable evaluators to examine changes in implementation and growth in student achievement over time, providing a much stronger test of the program's effects than what is typically obtained through a cross-sectional comparison.

In short, our evaluation provides some evidence that the instructional practices promoted by the SI programs have led to improved academic achievement, but the relationships are weak and our conclusions are affected by the assessment and design issues discussed above. Changes to data systems and large-scale assessment programs are needed to provide a means of conducting program evaluations that can provide clear evidence concerning the link between classroom practices and student achievement.

References

American Association for the Advancement of Science (1993). Benchmarks for science literacy: Project 2061. New York: Oxford University Press.

Cohen, D. K., & Ball, D. L. (1990). Relations between policy and practice: A commentary. Educational Evaluation and Policy Analysis, 12, 331-338.

Cohen, D. K., & Hill, H. C. (1998). State policy and classroom performance: Mathematics reform in California (CPRE Policy Brief). Philadelphia: Consortium for Policy Research in Education.

Consortium for Policy Research in Education (1995). Reforming science, mathematics, and technology education: NSF's State Systemic Initiatives (CPRE Policy Brief). New Brunswick, NJ: Author.

Corcoran, T. B., Shields, P. M., and Zucker, A. A. (1998). Evaluation of the National Science Foundation's Statewide Systemic Initiatives (SSI) Program: The SSI's and professional development for teachers. Menlo Park: SRI International.

Education Week (2000, January). Quality counts 2000: Who should teach? Bethesda, MD: Author.

Fantuzzo, J. W., King, J. A., & Heller, L. R. (1992). Effects of reciprocal peer tutoring on mathematics and school adjustment: A component analysis. Journal of Educational Psychology, 84, 331-339.

Fox, J. (1998). NSF programs attacked as weak, unclear. Education Daily, July 24, pp. 1-2.

Ginsburg-Block, M. D., & Fantuzzo, J. W. (1998). An evaluation of the relative effectiveness of NCTM standards-based interventions for low-achieving urban elementary students. Journal of Educational Psychology, 90, 560-569.

Goertz, M. E., & Duffy, M. C. (2001). Assessment and accountability systems in the 50 states: 1999-2000. Philadelphia: Consortium for Policy Research in Education.

Goldstein, H. (1995). Multilevel Statistical Models (2nd ed.). London: Arnold.

Greenwood, C. R., Carta, J. J., & Hall, R. V. (1988). The use of peer tutoring strategies in classroom management and educational instruction. School Psychology Review, 17, 258-275.

Hill, P. T. (1995). Reinventing public education. Santa Monica, CA: RAND.

Knapp, M. S. (1997). Between systemic reforms and the mathematics and science classroom: The dynamics of innovation, implementation, and professional learning. Review of Educational Research, 67, 227-266.

Koretz, D., and Barron, S. I. (1998, in press). The validity of gains on the Kentucky Instructional Results Information System (KIRIS). Santa Monica: RAND.

Linn, R. L. (2000). Assessments and accountability. Educational Researcher, 29 (2), 4-16.

Laguarda, K. G. (1998). Assessing the SSIs' impacts on student achievement: An imperfect science. Menlo Park, CA: SRI International.

Mayer, D. P. (1998). Do new teaching standards undermine performance on old tests? Educational Evaluation and Policy Analysis, 20, 53-73.

McCaffrey, D., & Bell, R. (1997). Bias reduction in standard error estimates for regression analyses from multi-stage designs with few primary sampling units. Paper presented at the Joint Statistical Meetings, Anaheim CA.

National Council of Teachers of Mathematics (1989). Curriculum and evaluation standards for school mathematics. Reston, VA: Author.

National Research Council (1996). National science education standards. Washington, DC: National Academy Press.

Schafer, J.L., (1997). Imputation of missing covariates under a general linear mixed model. Technical report available at http://www.stat.psu.edu/~jls/.

Shields, P. M., Corcoran, T. B., & Zucker, A. A. (1994). Evaluation of NSF's Statewide Systemic Initiatives (SSI) program: First-year report. Menlo Park, CA: SRI International.

Shields, P. M., Marsh, J. A., and Adelman, N. E. (1998). Evaluation of the National Science Foundation's Statewide Systemic Initiatives (SSI) Program: The SSI's Impacts on Classroom Practice. Menlo Park: SRI International.

Smith, M., & O'Day, J. (1991). Systemic school reform. In S. H. Fuhrman & B. Malen (Eds.), The politics of curriculum and testing (pp. 233-268). Bristol, PA: The Falmer Press.

Smerdon, B. A., Burkam, D. T., & Lee, V. E. (1999). Access to constructivist and didactic teaching: Who gets it? Where is it practiced? Teachers College Record, 101, 5-34.

Tyack, D., & Cuban, L. (1995). Tinkering toward utopia. Cambridge, MA: Harvard University Press.

Verschaffel, L., & De Corte, E. (1997). Teaching realistic mathematical modeling in the elementary school: A teaching experiment with fifth-graders. Journal for Research in Mathematics Education, 28, 577-601.

Webb, N. M., & Palincsar, A. S. (1996). Group processes in the classroom. In D. C. Berliner & R. C. Calfee (Eds.), Handbook of educational psychology (pp. 841-873). New York: Macmillan.

Weiss, I. R., Montgomery, D. L., Ridgway, C. J., & Bond, S. L. (1998). Local Systemic Change through Teacher Enhancement: Year three cross-site report. Chapel Hill, NC: Horizon Research, Inc.

Williams, L. (1998). The Urban Systemic Initiatives (USI) program of the National Science Foundation: Summary update. Washington, DC: NSF.

Figure 1

Coefficients for Reform Practices, Open-Response Math

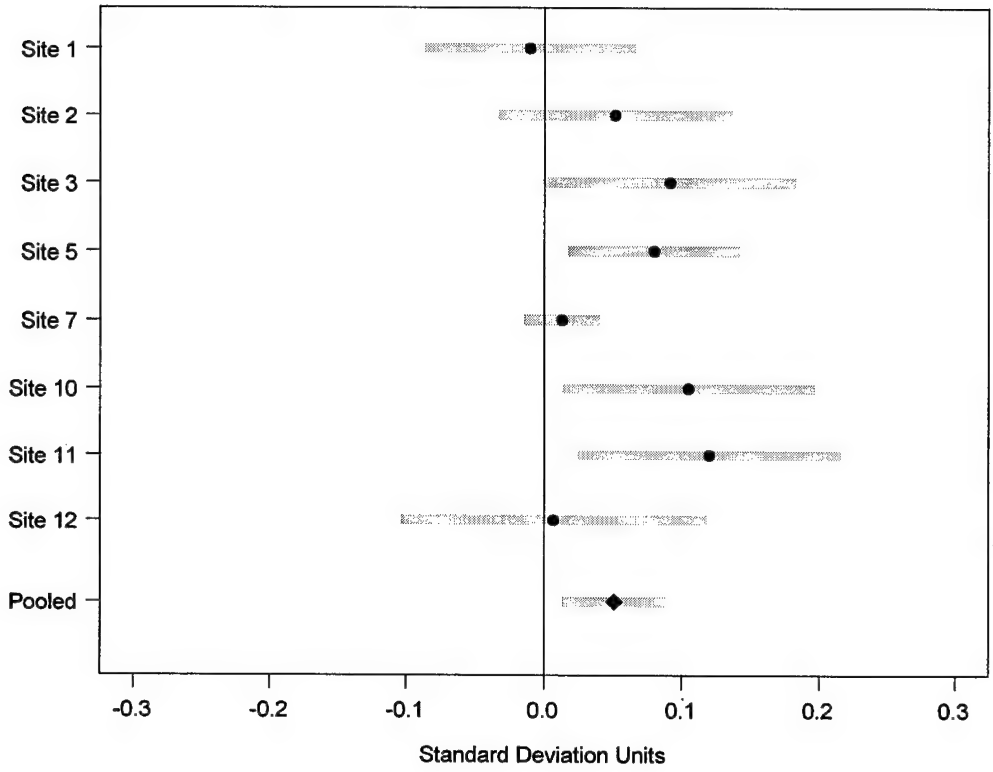


Figure 2

Coefficients for Reform Practices, Multiple-Choice Math

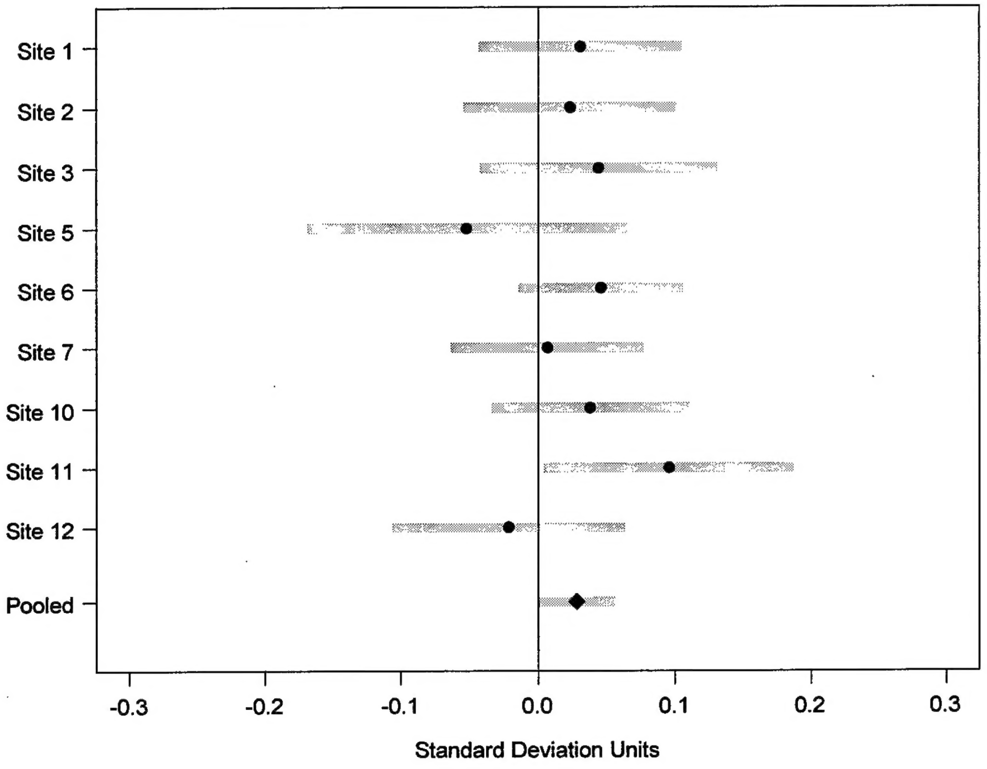


Figure 3

Coefficients for Reform Practices, Open-Response Science

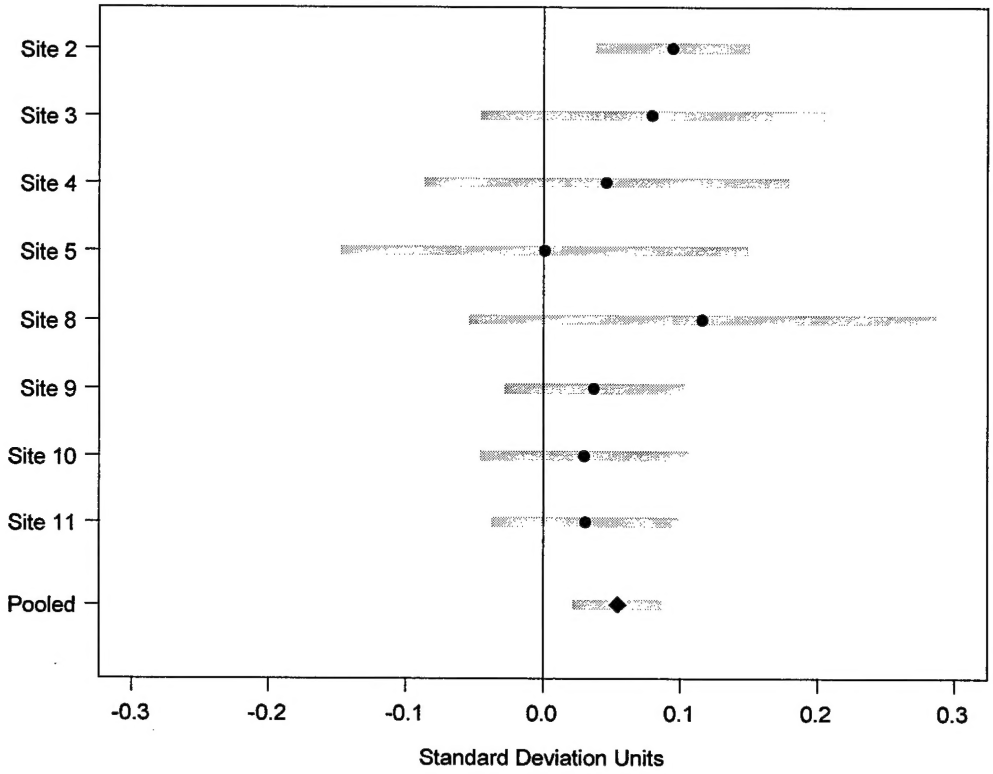
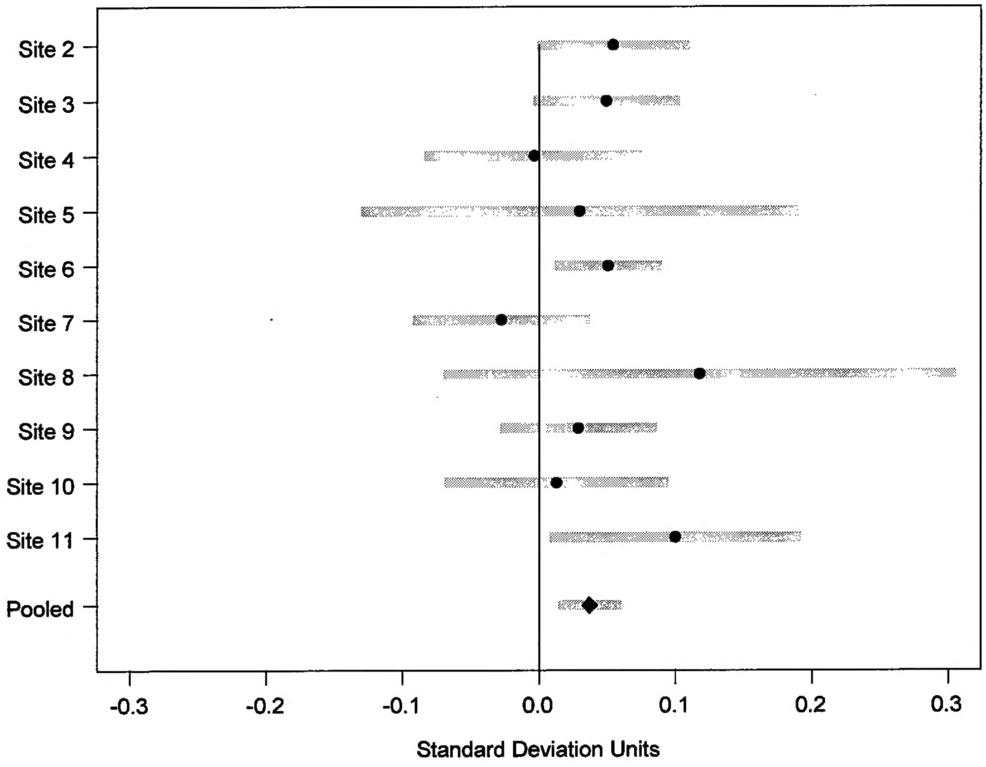


Figure 4

Coefficients for Reform Practices, Multiple-Choice Science



Notes

[1] As we discuss later, the lower reliability of the traditional practices scale will tend to attenuate relationships with achievement.

[2] In each model we tested whether a nonlinear term provided a better fit than a linear term. In all but two models that used the traditional practices scale, the linear term provided an adequate fit.

[3] We used our model to predict the score for the "average" student (a student with all student background predictors set to the mean) with a teacher scoring 3 on each reform practices item (monthly use of reform practices). We then found the percentile of this predicted score among the test scores from the site, and repeated the process for the average student with a teacher scoring 4 on each item (daily use). The percentile is based on our sample and is not a percentile from a national norming group.